In the Claims:

This section sets forth a clean version of the entire set of pending claim(s) under 37 C.F.R. 1.121(c)(3). Appendix A submitted herewith sets forth a marked up version of the prior pending claim(s) which have been amended by this Amendment with additions shown with underlining (e.g. new text) and deletions shown with a strikethrough (e.g. ~~delete text~~) under 37 C.F.R. 1.121(b)(1)(iii).

This amendment amends claims 1, 7-23, 27-29, 34, 39 and 41-42, adds claims 43-49, and cancels claims 2-6 and 40, without prejudice.

Sub B1 A1

1. A method for quantitatively representing documents in a vector space, comprising the steps of:

identifying a first document to be processed from a plurality of documents;

extracting a first feature corresponding to the first document from the plurality of documents, the first feature comprising text surrounding an image included in the document;

converting the first feature to a first vector; and

associating the first vector with the first document.

A2

7. The method of claim 1 further comprising the steps of:

extracting a second feature corresponding to the document, the second feature comprising a first URL representing the first document;

converting the second feature to a second vector; and

associating the second vector with the first document.

8. The method of claim 7, wherein the step of converting the second feature comprises the sub-steps of:

identifying each unique word within the URLs representing all documents in the collection of documents; and

counting the occurrences of each unique word in the first URL;

creating a vector having a number of dimensions equal to the number of unique words in the URLs representing all documents in the collection of documents, and further having as each element a numeric value representative of the number of occurrences in the first URL of the corresponding word.

A2

9. The method of claim 8, wherein the value representative of the number of occurrences in the first URL of the corresponding word is calculated as the token frequency weight of the corresponding word multiplied by the inverse context frequency weight of the corresponding word.

10. The method of claim 1 further comprising the steps of:

extracting a second feature corresponding to the first document, the second feature comprising inlinks in the collection of documents linking to the first document;

converting the second feature to a second vector; and

associating the second vector with the first document .

11. The method of claim 10, wherein the step of converting the second feature comprises the sub-steps of:

identifying each document having links within the collection of documents;

determining how many times each document having links points to the first document; and

creating the second vector having a number of dimensions equal to the number of documents having links in the collection of documents, and the second vector further having as each element a numeric value representative of the number of links in each corresponding document linking to the first document.

12. The method of claim 11, wherein the numeric value representative of the number of links in each corresponding document linking to the first document is calculated as

the token frequency weight of the corresponding link multiplied by the inverse context frequency weight of the corresponding link.

A2

13. The method of claim 10, wherein the step of converting the second feature comprises the sub-steps of:

identifying each document having hyperlinks within the collection of documents, and further identifying each unique word associated with URLs defining hyperlinks in each document;

counting the occurrences of each unique word in the URLs defining hyperlinks pointing to the first document; and

creating the second vector having a number of dimensions equal to the number of unique words associated with URLs defining hyperlinks within the collection of documents, and the second vector further having as each element a numeric value representative of the number of occurrences in the URLs defining hyperlinks pointing to the first document of the corresponding word.

14. The method of claim 13, wherein the numeric value representative of the number of occurrences in the URLs defining hyperlinks pointing to the first document of the corresponding word is calculated as the token frequency weight of the corresponding word multiplied by the inverse context frequency weight of the corresponding word.

15. The method of claim 1 further comprising the steps of:

extracting a second feature corresponding to the first document, the second feature comprising outlinks in the collection of documents linking to the first document;

converting the second feature to a second vector; and

associating the second vector with the first document .

16. The method of claim 15, wherein the step of converting the second feature comprises the sub-steps of:

identifying each other document linked to by all documents within the collection of documents; and

creating the second vector having a number of dimensions equal to the number of other documents linked to by documents in the collection of documents, and the second vector further having as each element a numeric value representative of the number of links in the first document linking to each corresponding other document.

A2

17. The method of claim 16, wherein the numeric value representative of the number of links in the first document linking to each corresponding other document is calculated as the token frequency weight of the corresponding link multiplied by the inverse context frequency weight of the corresponding link.

18. The method of claim 15, wherein the step of converting the second feature comprises the sub-steps of:

identifying each unique word associated with URLs defining hyperlinks in each document in the collection of documents;

counting the occurrences of each unique word in the URLs defining hyperlinks in the first document; and

creating the second vector having a number of dimensions equal to the number of unique words associated with the URLs defining hyperlinks in each document, and the second vector further having as each element a numeric value representative of the number of occurrences in the URLs defining hyperlinks in the first document of the corresponding word.

19. The method of claim 18, wherein the numeric value representative of the number of occurrences in the URLs defining hyperlinks in the first document of the corresponding word is calculated as the token frequency weight of the corresponding word multiplied by the inverse context frequency weight of the corresponding word.

20. The method of claim 49, wherein the second feature comprises a text genre feature.

21. The method of claim 20, wherein the step of converting the second feature comprises the sub-steps of:

for each possible text genre, processing the first document to calculate the probability that the first document is of the corresponding text genre; and

A2

creating the second vector having a number of dimensions equal to the number of possible text genres, and the second vector further having as each element a numeric value representative of the probability that the first document is of the corresponding genre.

22. The method of claim 49, wherein the first feature comprises the color histogram for an image represented by the first document.

23. The method of claim 22, wherein the step of converting the first feature comprises the sub-steps of:

quantizing the image represented by the first document into a multi-dimensional color model;

creating a color histogram having a plurality of bins for each dimension in the color model, each bin corresponding to a unique combination of binary bits representing information from the associated dimension of the color model;

counting each of a plurality of pixels from the image in a corresponding bin associated with each dimension of the color model; and

creating the first vector having a number of dimensions equal to the total number of bins in the color histogram, and the first vector further having as each element a numeric value representative of the number of pixels in the image corresponding to the corresponding histogram bin.

24. The method of claim 23, wherein the plurality of pixels from the image in the counting step comprises all of the pixels in the image.

25. The method of claim 24, wherein the plurality of pixels from the image in the counting step comprises an approximately uniformly spaced set of subsampled pixels from the image.

A2

26. The method of claim 23, wherein:

the color model comprises a three-dimensional hue, saturation, and value color model;

each dimension of the color model is represented by two bits of information; and

the color histogram has four bins for each dimension in the color model, for a total of twelve bins.

27. The method of claim 23, wherein the image represented by the first document comprises a region of a bitmap.

28. The method of claim 49, wherein the first feature comprises the color complexity of an image represented by the first document.

29. The method of claim 28, wherein the step of converting the first feature comprises the sub-steps of:

quantizing the image represented by the first document into a multi-dimensional color model;

determining the maximum number of pixels in any row in any image represented by any document in the collection of documents;

determining the maximum number of pixels in any column in any image represented by any document in the collection of documents;

creating a horizontal complexity histogram and a vertical complexity histogram, each having a number of bins equal to the maximum number of pixels in any row and in any column, respectively;

identifying horizontal runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of rows of the quantized image belonging to the horizontal runs in a corresponding bin of the horizontal complexity histogram;

A2

identifying vertical runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of columns of the quantized image belonging to the vertical runs in a corresponding bin of the horizontal complexity histogram;

creating a horizontal complexity vector having a number of dimensions equal to the maximum number of pixels in any row, and further having as each element a numeric value representing the number of pixels in the image in the corresponding horizontal histogram bin; and

creating a vertical complexity vector having a number of dimensions equal to the maximum number of pixels in any column, and further having as each element a numeric value representing the number of pixels in the image in the corresponding vertical histogram bin.

30. The method of claim 29, wherein the plurality of rows comprises all rows of the quantized image, and wherein the plurality of columns comprises all columns of the quantized image.

31. The method of claim 29, wherein the plurality of rows comprises an approximately uniformly spaced set of subsampled rows from the image, and wherein the plurality of columns comprises an approximately uniformly spaced set of subsampled columns from the image.

32. The method of claim 29, wherein:

the color model comprises a three-dimensional hue, saturation, and value color model; and

each dimension of the color model is represented by two bits of information.

33. The method of claim 29, further comprising the step of concatenating the horizontal complexity vector and the vertical complexity vector to form a complexity vector having a number of dimensions equal to the maximum number of pixels in any row plus the maximum number of pixels in any column.

A2

34. The method of claim 28, wherein the step of converting the first feature comprises the sub-steps of:

quantizing the image represented by the first document into a multi-dimensional color model;

determining the maximum number of pixels in any row in any image represented by any document in the collection of documents;

determining the maximum number of pixels in any column in any image represented by any document in the collection of documents;

creating a horizontal complexity histogram and a vertical complexity histogram, each having a selected number of bins corresponding to a plurality of quantized ranges of run lengths;

identifying horizontal runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of rows of the quantized image belonging to the horizontal runs in a corresponding bin of the horizontal complexity histogram;

identifying vertical runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of columns of the quantized image belonging to the vertical runs in a corresponding bin of the horizontal complexity histogram;

creating a horizontal complexity vector having a number of dimensions equal to the selected number of bins in the horizontal complexity histogram, and further having as each element a numeric value representing the number of pixels in the image in the corresponding horizontal histogram bin; and

creating a vertical complexity vector having a number of dimensions equal to the number of bins in the vertical complexity histogram, and further having as each element a numeric value representing the number of pixels in the image in the corresponding vertical histogram bin.

A2

35. The method of claim 34, wherein:

a bin b_x in the horizontal complexity histogram corresponding to a horizontal run of length r_x is identified by a relationship $b_x = \text{floor}(r_x(N\text{-}1) / (n_x/4)) + 1$, where N is the selected number of bins in the horizontal complexity histogram and n_x is a maximum number of pixels in any row of an image in the collection; and

a bin b_y in the vertical complexity histogram corresponding to a vertical run of length r_y is identified by a relationship $b_y = \text{floor}(r_y(N\text{-}1) / (n_y/4)) + 1$, where N is the selected number of bins in the horizontal complexity histogram and n_y is a maximum number of pixels in any row of an image in the collection.

36. The method of claim 34, wherein the plurality of rows comprises an approximately uniformly spaced set of subsampled rows from the image, and wherein the plurality of columns comprises an approximately uniformly spaced set of subsampled columns from the image.

37. The method of claim 34, wherein:

the color model comprises a three-dimensional hue, saturation, and value color model; and

each dimension of the color model is represented by two bits of information.

38. The method of claim 34, further comprising the step of concatenating the horizontal complexity vector and the vertical complexity vector to form a complexity vector having a number of dimensions equal to the selected number of bins in the

horizontal complexity histogram plus the selected number of bins in the vertical complexity histogram.

39. A method for quantitatively representing in a vector space users of a collection of documents, comprising the steps of:

identifying a first user to be processed from the users of the collection of documents;

extracting from the collection of documents a first feature representing a first sub-set of documents of the collection that have been accessed by the first user;

A2

converting the first feature to a first vector; and

associating the first vector with the first user.

41. The method of claim 39, wherein the converting step comprises the steps of:

identifying each unique document in the collection of documents;

calculating the number of times the first user accessed each document in the collection of documents; and

creating the first vector having a number of dimensions equal to the number of documents in the collection of documents, and the first vector further having as each element a numeric value representative of the number of times the first user has accessed the corresponding document.

42. The method of claim 41, wherein the value representative of the number of times the first user has accessed the corresponding document is calculated as the token frequency weight of the corresponding document multiplied by the inverse context frequency weight of the corresponding document.

43. A computer-readable medium containing instructions for causing a computer-system to quantitatively representing documents in a vector space, by the steps of:

A3

identifying a document to be processed from a plurality of documents;

selecting a first feature from a set of multi-modal features including a text feature, a hyperlink feature, an image feature, a user information feature, and a genre feature;

extracting from the document information associated with the first feature;

converting information associated with the first feature into a first vector;

associating the first vector with the document;

selecting a second feature from the set of multi-modal features;

extracting from the document information associated with the second feature;

converting the information associated with the second feature into a second vector; and

associating the second vector with the document.

A3

44. The computer-readable medium of claim 43 wherein the first feature is an image feature.

45. The computer-readable medium of claim 44 wherein the first feature comprises a color histogram for an image included in the document.

46. The computer-readable medium of claim 45 wherein converting the information associated with the first feature into the first vector comprises the steps of:

quantizing the image included in the document into a multi-dimensional color model;

creating a color histogram having a plurality of bins for each dimension in the color model, each bin corresponding to a unique combination of binary bits representing information from the associated dimension of the color model;

counting each of a plurality of pixels from the image in a corresponding bin associated with each dimension of the color model; and

creating a vector having a number of dimensions equal to the total number of bins in the color histogram, and further having as each element a numeric value representative of the number of pixels in the image corresponding to the corresponding histogram bin.

A3

47. The computer-readable medium of claim 44 wherein the first feature comprises color complexity of an image comprising part of the document.

48. The computer-readable medium of claim 47 wherein converting the information associated with the first feature into the first vector comprises the steps of:

quantizing the image included in the document into a multi-dimensional color model;

determining the maximum number of pixels in any row in any image represented by any document in the collection of documents;

determining the maximum number of pixels in any column in any image represented by any document in the collection of documents;

creating a horizontal complexity histogram and a vertical complexity histogram, each having a number of bins equal to the maximum number of pixels in any row and in any column, respectively;

identifying horizontal runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of rows of the quantized image belonging to the horizontal runs in a corresponding bin of the horizontal complexity histogram;

identifying vertical runs of pixels of all possible lengths in the quantized image, and for each possible length, counting the number of pixels in a plurality of columns of the quantized image belonging to the vertical runs in a corresponding bin of the horizontal complexity histogram;

creating a horizontal complexity vector having a number of dimensions equal to the maximum number of pixels in any row, and further having as each element a numeric value representing the number of pixels in the image in the corresponding horizontal histogram bin; and creating a vertical complexity vector having a number of dimensions equal to the maximum number of pixels in any column, and further having as each element a numeric value representing the number of pixels in the image in the corresponding vertical histogram bin.

A3

49. A method for quantitatively representing documents in a vector space, comprising the steps of:

identifying a first document to be processed from a plurality of documents;

extracting a first feature corresponding to the first document from the plurality of documents, the first feature comprising an image feature;

converting the first feature to a first vector;

associating the first vector with the first document;

extracting a second feature corresponding to the document, the second feature comprising a one of a text feature, a hyperlink feature, a user feature and a genre feature;

converting the second feature into a second vector; and

associating the second vector with the first document.